Quantitative approaches: modelling variation and change in Occitan

Damien Mooney
University of Bristol

1. Introduction

The ‘variable rule’ has been a central theoretical construct in variationist sociolinguistics since Labov (1969) first introduced it in his analysis of African-American Vernacular English (AAVE) copula contraction and deletion. This construct has as its basis the notion of ‘orderly heterogeneity’ (Weinreich/Labov/Herzog 1968, 100), or the postulate that language variation and language change are constrained by a combination of (potentially interacting) social and linguistic factors: “language is not random or free, but systematic and rule governed” (Tagliamonte 2006, 129). Variable rules are ‘abstract optional rules’ which form an integral part of a language variety’s structural description: “once accepted and incorporated into description, variability can be made a function not only of the presence or absence of linguistic elements, but also can be constrained by extralinguistic factors, all within the same notional and theoretical framework” (Cedergren/Sankoff 1974, 333–334). The extent to which variable rules reflect actual linguistic competence at the level of the individual and of the ‘speech community’ has been a matter of some theoretical debate (see, for example, Sankoff/Labov 1979); from the linguist’s perspective, however, the variable rule can be considered as “the probabilistic modelling and statistical treatment of discrete choices and their conditioning” (Sankoff 1988, 984). This chapter presents and evaluates different techniques, in the variationist sociolinguist’s toolkit, that can be used to undertake this statistical treatment, illustrating these quantitative methods with sociophonetic data from Occitan.

During the 1970s, variable rule analysis was further developed in studies of language variation and change and this included the development of the ‘variable rule program’ (Cedergren/Sankoff 1974; Rousseau/Sankoff 1978), a statistical modelling package for sociolinguistic analysis which provided a means of estimating the parameters of variable rules (Johnson 2009, 359). The variable rule program was created as a response to the fact that existing statistical modelling techniques, such as 'analysis of variance' ANOVA, were largely unsuitable for analysing spontaneous speech data, which are notoriously unbalanced in their distribution: “in order to model a grammar that has heterogeneity with contextually conditioned 'order' to it as well as innumerable black regions, a mathematical construct had to

be devised that would suitably mirror it" (Tagliamonte 2006, 130). The variable rule program has existed under many guises since its initial development (Tagliamonte/Baayen 2012: 136): *Varbrul* (Cedergren/Sankoff 1974); *Goldvarb* 2.0 (Rand/Sankoff 1990); *Goldvarb X* (Sankoff 2005); *Goldvarb Lion* (Sankoff/Tagliamonte/Smith 2012). These packages, often collectively referred to as VARBRUL, have allowed the sociolinguist to model statistically the distribution of two discrete linguistic variants, as well as the (collective) effect of social and linguistic factors that condition the variation observed. Tagliamonte/Baayen note, however, that the past 30 years have seen the development of more sophisticated statistical modelling techniques, which may be more appropriate for analysing language data: "currently, the debate centers not on whether statistical methods are appropriate, but on the choice of which one is best" (2012, 136). Packages such as *Rvarb* (Paolillo 2002), *Rbrul* (Johnson 2009) and *R* (R Development Core Team 2009) provide the analyst with the opportunity to implement a more advanced version of variable rule analysis; these approaches have the primary advantage of facilitating a higher level of generalisability to the wider population with respect to the results obtained.

Traditionally, structured variability in Romance varieties has received relatively little attention when compared with the large body of variationist sociolinguistic literature on variable rule analysis in varieties of English. This is perhaps a consequence of a wider tendency, in European dialectological studies, to focus on 'lexical isoglosses, lexical incidence, or unconnected phonetic variables' (Labov 2007, 348); studies of the Romance languages have tended to examine low-level phonetic transfer and change, rather than investigating the social and linguistic constraints that govern these developments. There are some variationist studies, however, which have presented applications of the (traditional) variable rule program to varieties of French such as, for example, Ashby (1981) and Van Compernolle (2008) on negative particle deletion, Ashby (1982; 1988) on left- and right-dislocations, Ashby (1992) and Williams/Van Compernolle (2009) on forms of address; Regan (1996) on second language acquisition, Moisset (2000) on variable liaison, and Temple (2000a; 2000b) on plosive devoicing. Variationist studies of Canadian varieties of French has been particularly progressive in using the variable rule program to analyse spontaneous speech data; for example, Paradis/Deshaies (1990) on stress alignment in Québec, Poplack (1992) on the subjunctive, Nagy/Blondeau (1999) on double subject marking in Montréal, King/Nadasdi (2003) on future temporal reference in Acadia, Sankoff and Blondeau (2007) on rhotics in Montréal, King/Martineau/Mougeon (2011) on first-person plural pronouns, and Comeau/King/Butler (2012) on past-tense aspectual distinctions in Acadia. More recently, other researchers have taken advantage of the advanced modelling techniques offered by the

R environment such as, for example, Roberts (2012) on future temporal reference in Martinique, Burnett/Tremblay/Blondeau (2015) on negative concord in Montréal, and Mooney (2015b; 2016a; 2016b) on dialect levelling in the phonological system of southwestern metropolitan French. To my knowledge, no studies of language variation and change have used the variable rule program on data from minority languages in the francophone context, the so-called *langues de France*, or regional languages. Some researchers have performed ANOVAs on France's regional languages, such as Villeneuve/Auger (2013) on subject-doubling and negative particle deletion in Picard, Sichel-Bazin/Buthke/Meisenburg (2012) on Occitan prosody, and Kennard and Lahiri (2015a; 2015b) on mutation in Breton; with the exception of Villeneuve/Auger (2013), the data presented in these studies was largely experimental and therefore more suited to ANOVA than studies of spontaneous speech (Tagliamonte 2006, 130).

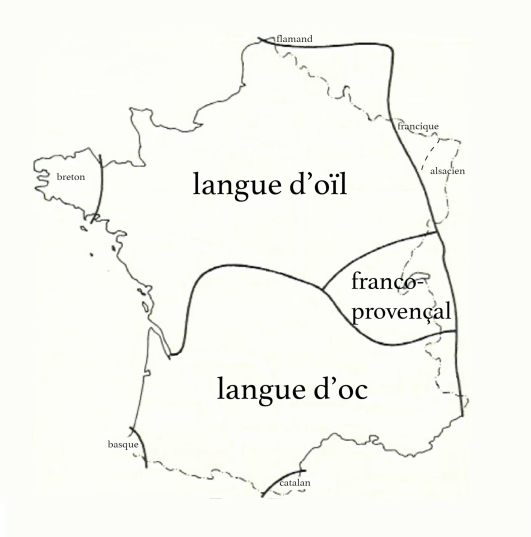

Figure 1. Gallo-Romance languages (Mooney 2016b, 9)

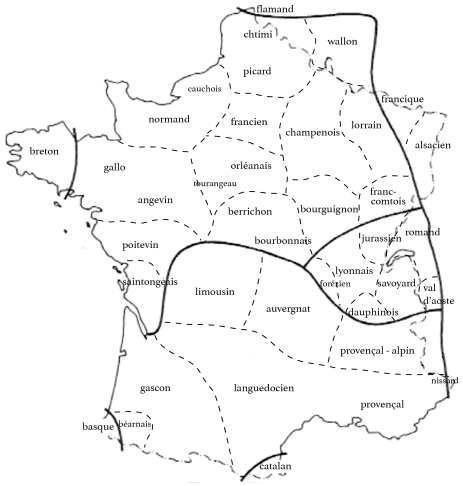

Figure 2. Gallo-Romance dialects (Mooney, 2016b, 9)

The statistical analyses presented in this chapter model linguistic variation and change in the consonantal and vocalic systems of a local variety of the 'Occitan' language. The most significant division within Gallo-Romance is between between the dialect area in the north, the *langue d'oïl*, and the dialect area in the south, known as the *langue d'oc* (see Figure 1). The modern *langue d'oc* area is commonly divided into six main dialectal areas (see Figure 2): *gascon* in the southwest, including the *béarnais* and *aranais* sub-dialects; central *languedocien*; *limousin* and *auvergnat* in the north; *provençal* in the southeast, including the *nissart* sub-dialect; *vivaro-alpin* or *provençal alpin* above the *provençal* region. In the second half of the twentieth century, the establishment of the *Institut d'Études Occitanes* led to the

use of the term 'Occitan' to refer to all *langue d'oc* dialects, collectively considered to be a single language. The use of the term 'Occitan' has become a source of ideological conflict in southern France, especially for those who consider local varieties of the *langue d'oc* to be languages in their own right (see Mooney 2015a; Moreux 2004; Blanchet/Schiffman 2004 for discussion); nonetheless, I will use the term 'Occitan' here for simplicity. The data presented in this chapter comes from the *béarnais* sub-dialect of *gascon*, spoken in the region of Béarn or the historically Romance-speaking part of the Pyrénées-Atlantiques *département* in southwestern France. *Gascon* is often singled out as a special case in Occitan dialectology because it contains strongly marked regional phonological and morphosyntactic features that are not found in any other dialects (Walter 1988, 153). Like all Occitan dialects, *béarnais* has found itself in an increasing state of language obsolescence from the late nineteenth century onwards. In the entire *gascon* region, the highest concentration of speakers exists in Béarn, making *béarnais* the principal surviving dialect. Moreux (2004) suggests that, at the beginning of the twentieth century, there were about 40,000 fluent native speakers of Occitan in Béarn, noting that the large majority of these speakers were over the age of 65 and rural dwellers.

This chapter begins by describing the Occitan data set collected in Béarn, and by outlining the linguistic variables to be modelled statistically – the dependent variables (Section 2): (i) rhotic consonants; (ii) front mid-vowel contrast. The social and linguistic factors expected to condition variation and change in the Occitan phonological inventory are then presented – the independent variables (Section 3), before an interim discussion of the methods used to normalise sociophonetic data for vocalic variables (Section 4), a necessary step which must be taken before reliable statistical analyses can take place. The body of the chapter presents a series of (increasingly complex) statistical modelling techniques for the Occitan linguistic variables under consideration, beginning with a traditional VARBRUL-style analysis of the Occitan rhotics (Section 5.1), before discussing interactions between independent variables and some methodological issues involved including correlated social and/or linguistic factors in the analyses proposed (Section 5.2). The front-mid vowels are submitted to an *Rbrul* analysis in Section 5.3, using a technique previously unavailable in the VARBRUL suite. The Occitan data is then submitted to a series of the most up-to-date statistical modelling techniques available for variationist data (Section 5.4), with the final section discussing some proposed statistical techniques for resolving on-going issues encountered with current modelling methodologies.

2. Dependent variables: Occitan sociophonetic data

In variationist sociolinguistics, the 'dependent variable' is the linguistic variable whose distribution we are interested in analysing statistically. Dependent variables in sociolinguistic studies are usually either binary or continuous: binary variables have two discrete variants and are categorical in nature; continuous variables are characterised by having a range of variants on a gradient scale (Hay 2011, 200). The Occitan data presented in this section is sociophonetic in nature, meaning that it was collected using traditional Labovian sociolinguistic methods and that it was analysed using the acoustic phonetic techniques of laboratory phonetics: "in sociophonetic research, the dependent variable is usually a phonetic factor, or a response in an experiment" (Hay 2011, 200). The data set contains examples of both binary and continuous dependent variables, the rhotic consonants and the front mid-vowels, respectively. The Occitan corpus contains high quality acoustic data, collected in 2012, for ten bilingual Occitan-French speakers, five male and five female, over the age of 65, and native to the region of Béarn. All informants participated in a wordlist translation task from French into Occitan and were recorded using a sampling rate of 44.1 kHz and a 16-bit PCM sample size on a Marantz PMD661 Solid State Sound Recorder. Subsequent acoustic analyses were performed in Praat version 5.2.21 (Boersma 2001; Boersma/Weenink 2012).

2.1 Categorical variables: Occitan rhotic consonants

There are very few comprehensive analyses of the distribution of rhotic consonants in Occitan varieties and so we are limited to what are often cursory observations based on small data sets. Most commentators agree, however, that *gascon* has two historically appropriate rhotic consonants, the voiced apical trill [r], and the voice apical tap [ɾ], which are in contrastive distribution in intervocalic position, e.g. *poret* /puˈret/ (‘chicken’) ~ *porret* /puˈret/ (‘leek’), (Bec 1973; Cardaillac Kelly 1973), and not contrastive in other contexts, such that "an archiphoneme could be set up for all other positions" (Cardaillac Kelly 1973, 32). The apical rhotics are not, however, in strictly complementary distribution in non-intervocalic contexts: the distribution of [r] and [ɾ] is somewhat constrained by their position within the syllable and with respect to word boundaries with a tendency for [r] to occur word-initially and as an onset after [n], and [ɾ] to occur in onset clusters and in the syllable coda, but this distribution is by

no means categorical (Cardaillac Kelly 1973, 32; Mooney 2014, 345).

Previous analyses of the *gascon* rhotics have noted the transfer of dorsal rhotic consonants from French due to prolonged language contact. The phonological inventory of modern standard French contains one rhotic consonant phoneme, the voiced uvular fricative /ʁ/, which is often realised as a voiced uvular trill [ʀ] by older rural speakers. Additionally, there is evidence to suggest that /ʁ/ is being realised increasing as a voiced uvular approximant, particularly in final position (Fougeron and Smith, 1999: 80). Cardaillac Kelly found that dorsal realisations [ʁ ʀ] occurred as variants of /r/ and /ɾ/ "in *all* positions as a consequence of bilingualism" (1973, 32), that when dorsal variants are used in intervocalic position, the phonemic distinction between /ɾ/ and /r/ is neutralised, e.g. *poret* ~ *porret* [puˈʁet], and that female speakers used more dorsal variants than male speakers.

The analysis of the Occitan rhotic consonants considered contact-induced change of the place of articulation of the categorical dependent variable, (R), with binary variants [apical] and [dorsal], representing the historically appropriate and transferred forms, respectively. 466 tokens of the (R) variable were categorised on the basis of an auditory or impressionistic analysis, which was supplemented by visual inspection of the acoustic spectrogram; an equal number token counts were extracted for both male and female speakers.

2.2 Continuous variables: Occitan front mid-vowels

Traditionally, Occitan distinguishes between two mid-vowels, /e/ and /ɛ/, in the front of the vowel space (e.g., *peis* /peʃ/ 'fish', *pè* /pɛ/ 'foot'); these vowels are contrastive phonemes in *gascon*, e.g., *qu'ei* /kej/ 'he is', *qu'ai* /kɛj/ 'I have'. There is some evidence to suggest that this phonemic distinction is not maintained in certain varieties of Occitan (Séguy 1973); the analysis of the front mid-vowels aimed to determine the extent to which this contrast is maintained in Béarn. Figure 3 presents the full Occitan oral vowel system.

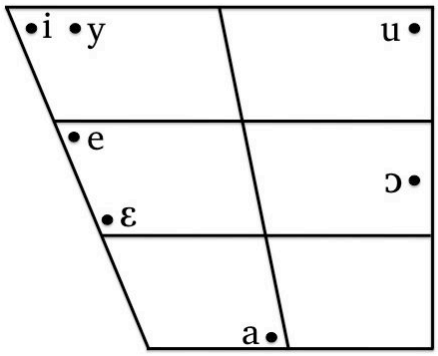

Figure 3. Occitan oral vowels (Mooney, 2014: 346)

While the phonemic distinction between the front mid-vowels is theoretically categorical, /e/ or /ɛ/, their phonetic realisations can be analysed as a continuous variable (E) by measuring the first two formant frequencies, F1 and F2, of each vowel token in the corpus; these formant values are located on a gradient scale in the acoustic vowel space. Formant frequencies are commonly held, in acoustic phonetic studies of oral vowels, to have general non-linear articulatory correlates: F1 exhibits an inverse correlation with vowel height; F2 exhibits a positive correlation with vowel frontness/backness. The first and second formants were estimated in Praat using the LPC (Linear Predictive Coding) algorithm, with a maximum of 4,000 Hz for male speakers and 4,500 Hz for female speakers. This instrumental adjustment based on biological sex was included as formant trackers may accurately track three formants below 4,500 Hz for female speakers, but may be less accurate for male speakers who might have four formants in the 4,500 Hz range (Clopper 2011, 195; Llamas/Watt/Johnson 2009, 392). The vowel onset and offset were first labelled in a Praat text grid and a script was used to automatically extract the value of F1 and F2 at the vowel midpoint; 253 tokens of the Occitan front mid-vowels were included in the analysis.

3. Independent variables: social and linguistic factors

Independent variables are generally linguistic or social factors that we expect to influence the distribution of the dependent variable: "features of the linguistic or extralinguistic context which impinge on the choice of one variant over the other" (Tagliamonte 2006, 131). In traditional variationist studies, independent variables are referred to as 'factor groups' and their variants are referred to as 'factors'. For example, an independent variable for speaker sex may be included in the analysis as a 'sex' factor group with variants [male] and [female].

These terms are not used outside the context of the variable rule program; in general statistical practice, factor groups are referred to as 'predictors' and factors are referred to as 'levels' Following Tagliamonte and Baayen, I use standard statistical terminology in the analyses and discussion that follow in order to "promote a bridge from sociolinguistic practice to 'general statistical practice'" (2012, 139).

For the categorical (R) variable, the analyses included five independent variables or predictors, two social and three linguistic (see Table 1). Speaker sex and the speaker's place of origin were included in the analysis to determine the extent to which the distribution of apical and dorsal variants was influenced by a speaker's gender and/or regional origin. Three linguistic predictors were also included; previous studies of the *gascon* rhotics have suggested that their distribution is partially constrained by syllable type and by phonological context. For the latter, a general distinction has been drawn between front and back consonants and between front and back vowels for the 'preceding phoneme' and 'following phoneme' predictors, the hypothesis being that adjacent anterior articulations will favour apical realisations and that posterior articulations will favour dorsal realisations. We must note, at this point, that the 'syllable type' predictor cannot be included in the same statistical analysis as 'preceding phoneme' and 'following phoneme' because these independent variables exhibit a collinear relationship in the data set in that they are highly correlated (see Section 5.2 for discussion). In the statistical analyses that follow, the models include either 'syllable type' alone or 'preceding phoneme' and 'following phoneme' together.

Predictor	Levels
Speaker sex	Male
	Female
Place	Gan
	Nousty
	Nay
Syllable type	Simple onset
	Complex onset
	Simple coda
	Complex coda
Preceding phoneme	Front vowel
	Back vowel
	Apical consonant
	Dorsal consonant
	Non-lingual consonant
	Pause
Following phoneme	Front vowel
	Back vowel
	Apical consonant
	Dorsal consonant
	Non-lingual consonant
	Pause

Table 1. *Independent variables included in statistical analyses of dependent variable (R)*

For the continuous (E) variable, the analyses included seven independent variables or predictors: two social and four linguistic. Again, speaker sex and place or origin were included as social predictors. 'Phoneme' was included as a predictor to determine the extent to which the historically appropriate phoneme (/e/ or /ɛ/) could predict F1 and F2 values when phonological context had been taken into account. Syllable type was also included in as an independent variable as the distribution of the front mid-vowels is heavily influenced by open and closed syllabic contexts in the local variety of French spoken in the region (Mooney, 2016b); as with the (R) variable, it was not possible to include 'syllable type' predictor in the same statistical analysis as the 'preceding phoneme' and 'following phoneme' predictors due to issues with collinearity (see Section 5.2). In the F1 statistical analyses, F2 was included as a predictor to investigate potential significant correlations between the formant frequencies; F1 was equally included as a predictor in the statistical models containing F2 as a dependent variable.

Predictor	Levels
Speaker sex	Male
	Female
Place	Gan
	Nousty
	Nay
Phoneme	/e/
	/ɛ/
Syllable type	/Cv#/– Open final
	/CvC#/– Closed final
	/vCV(C)#/– Open medial
Preceding phoneme	Various
Following phoneme	Various
F1 or F2	Continuous

Table 2. *Independent variables included in statistical analyses of dependent variable (E)*

4. Normalising vocalic data

Before undertaking a statistical analysis of the formant frequencies (F1 and F2) extracted from vocalic data, it is first necessary to normalise the data set. This is because different speakers exhibit variation in the formant values they produce for a given phonological vowel because of physiological differences in their vocal tracts. These differences are most notable when comparing speakers of different biological sexes because of the greater size of the male vocal tract and the lower fundamental frequency of the source sound. There exists, however, no constant difference between the varying dimensions of the male and female vocal tracts. In order to account for this sexual dimorphism, for the fact that no two speakers' vocal tracts share the same dimensions and to reliably compare vowel tokens across speakers and sexes, all data must be normalised. Normalisation aims to eliminate variation which is caused by anatomical differences while preserving variation that is sociolinguistically significant.

Normalisation procedures are categorised as follows (following Flynn 2011, 3): vowel-intrinsic or extrinsic; formant-intrinsic or extrinsic; speaker-intrinsic or extrinsic. Vowel-intrinsic procedures use information from a single vowel category to normalise a given token while vowel-extrinsic procedures use information from multiple vowel categories to normalise each token. Similarly, formant-intrinsic methods use information from one formant to normalise a given token. For example, all F1 values are taken into account when

normalising a given F1 token. On the other hand, formant-extrinsic procedures use multiple formants (F1, F2, F3, etc.) to normalise a single F1 value. Finally, speaker-intrinsic techniques use information from a single speaker to normalise their formant values, while speaker-extrinsic methods use information from all speakers in the sample population to normalise an individual's vowels. Comparative studies on the reliability of various normalisation techniques have shown that the most reliable procedures are classified as vowel-extrinsic, formant-intrinsic and speaker-intrinsic (Adank/Smits/van Hout 2004; Flynn 2011; Flynn/Foulkes 2011, 683). Thus, successful normalisation procedures use multiple vowel values within one formant for an individual speaker to normalise their data set.

Flynn and Foulkes (2011) provide an overview of twenty vowel normalisation methods, six vowel-intrinsic and fourteen vowel-extrinsic, many of which are available via the online normalisation tool NORM (Thomas/Kendall 2007). The earliest example of a vowel-extrinsic, formant-intrinsic, speaker-intrinsic normalisation procedure was proposed by Lobanov (1971), which expresses values relative to the hypothetical centre of a speaker's vowel space. While many newer normalisation procedures exist, such as the Watt and Fabricius (2002) method and the Nearey (1977) method, I have chosen to use the Lobanov method for the following reasons: Fabricius/Watt/Johnson (2009) showed that Lobanov's method outperformed both the Watt and Fabricius and the Nearey methods to a statistically significant extent; Adank/Smits/van Hout (2004) determined the Lobanov method to be the best for reducing the effect of anatomical variation while maintaining phonological and sociolinguistic variation; Flynn confirmed that 'its existing use in the sociolinguistic world is […] warranted' (2011, 22).

The Lobanov method involves calculating z-scores for individual data points which has the effect of transforming the original distribution to one in which the mean becomes zero and the standard deviation becomes 1. A z-score quantifies the original value relative to the number of standard deviations that the score is from the mean of the distribution. Z-scores (z) were obtained by first calculating the mean (μ) of F1 and F2 values separately (formant-intrinsic) across vowel types (vowel-extrinsic) for all of an individual speaker's data (speaker-intrinsic). The mean value of the relevant formant was then subtracted from each individual token (x) for F1 and F2. The result was then divided by the speaker's standard deviation (σ) of the mean frequency for the relevant formant (procedure described in Chen 2008, 635). The formula for z-score calculation can be summarised as: $z = (x - \mu)/\sigma$. The resultant z-scores are all centred around a mean of (0, 0) for each individual speaker, enabling comparison across speakers as the mean of all of their distributions is the same.

5. Statistical modelling

Statistical modelling allows the sociolinguist to identify the various components of a variable rule using statistical inference. There are many statistical tests that can be used to examine the effect of an independent variable on the distribution of the variants of a dependent variable, such as a t-test or Spearman's correlation (Hay 2011, 206–207), but these are largely inappropriate for sociolinguistic or sociophonetic data sets or, put simply, for the analysis of spontaneous speech data. This is because it is not possible to use these tests to consider the effect of multiple (potentially interacting) predictors on a dependent variable, the essence of a variable rule. In order to obtain "an assessment of the significance of each candidate predictor over and above any variation that can be explained by the other potential predictors" (Hay 2011, 207), we must use a statistical modelling technique known as regression:

> "Regression analysis is a statistical tool for the investigation of relationships between variables. […] To explore such issues, the investigator assembles the data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence" (Sykes 1993, 1).

ANOVA is a special case of regression that has been widely used in experimental linguistic studies, but this method is not suitable for spontaneous speech data as it assumes an even distribution of the data across the cells of the data set and this is almost never the case in sociolinguistic studies (Tagliamonte 2006, 137).

The regression models presented in this chapter have all been carried out in the R environment (version 3.2.3) using the *Rbrul* (version 2.3.2) text-based interface (Johnson 2008) which makes use of existing functions in the R environment, particularly the model-fitting fucntions *glm* and *glmer* (Bates/Sarkar 2008). The remainder of this section presents a series of increasingly complex regression analyses for categorical and continuous dependent and independent variables with the aim of bridging the gap between traditional variationist analyses using the variable rule program and more advanced general statistical practice. Where possible, the regression models have been presented as they appear in the *Rbrul* interface to familiarise the reader with R output. All models distinguish the follow levels for statistical significance: $p < .05$ and $p < .01$, for which the probability of observing the effect returned by chance is less than 5% and 1%, respectively; $p < .001$ is highly significant; for $p <$

.0001, the probability of observing the result returned is considered to be approximately zero ($p \approx 0$), or 100%.

5.1 Logistic regression models

The statistical modelling technique most widely used in sociolinguistics, and most suited to the analysis of categorical variables in spontaneous speech data, is logistic regression, a type of 'generalised linear model' (Agresti 2007, 67): logistic regression examines the effect of multiple predictors on a binary categorical dependent variable. The variable rule program, or VARBRUL analysis, is "an implementation of logistic regression for sociolinguists" (Hay 2011, 207) that allows the analyst to model the effect of categorical predictors ('factor groups') on a categorical variable; the versions of VARBRUL that are currently available do not support continuous dependent or independent variables.[1] The Occitan rhotic consonant data set is modelled statistically in this section using (i) a traditional VARBRUL-style binomial stepwise regression analysis, and (ii) a simple main effects one-level binomial regression analysis in *Rbrul*.

The VARBRUL series of applications offers two options for data analysis (Tagliamonte 2006, 139): (i) binomial one-step and (ii) binomial step-up/step-down. The first option provides statistical information on all predictors included in the analysis, including those that are not determined to have a significant effect ($p < .05$) on the dependent variable. The second option, also known as stepwise regression, has been used most often in studies of language variation and change (Tagliamonte 2006, 140):

> "Stepping up, it starts with no predictors and adds the most significant factor group, if there is one, before repeating the procedure. Stepping down, it starts with all possible predictors and removes the one that contributes least to the model, and then repeats this until all remaining predictors are significant" (Johnson 2009, 380).

When the stepwise procedure is complete, VARBRUL returns a logistic regression model that includes all 'factor groups', or predictors, that "affect the response variable of interest, in what direction and to what degree" (Johnson 2009, 359).

---

[1] *Varbrul 3* allows continuous variables to be included in the analysis but this version of VARBRUL analysis has not been made available on personal computers (Sankoff 2006, 1157; Johnson 2009, 217).

Factor group	*Factors*	*Total N*	*% of total*	*Factor weight*
Place of origin	Nay	145	32	0.708
	Gan	231	22	0.615
	Nousty	90	6	0.205
Following phoneme	Pause	66	55	0.817
	Apical consonant	77	27	0.569
	Non-lingual consonant	43	26	0.510
	Front vowel	214	12	0.402
	Dorsal consonant	9	11	0.348
	Back vowel	57	12	0.313
Preceding phoneme	Pause	17	41	0.869
	Front vowel	215	29	0.658
	Back vowel	89	28	0.560
	Dorsal consonant	50	10	0.476
	Non-lingual consonant	61	3	0.211
	Apical consonant	34	3	0.203

Table 3. VARBRUL-style stepwise regression analysis of Occitan rhotic consonants, with [dorsal] as application value (Log likelihood = –193.429, degrees of freedom = 12, significance = 0.000, input = 0.107)

Table 3 presents the results of a typical VARBRUL-style[2] stepwise logistic regression, using VARBRUL terminology for the categorical Occitan (R) dependent variable, with variants [dorsal] and [apical]. For binary categorical variables, one variant must be designated as the 'application' or 'response' value, or the "variant defined as the outcome of the variable rule" (Tagliamonte 2006, 263). The results returned by the variable rule program are relative to the application value. For example, the model presented in Table 3 included [dorsal] as the application value and so any significant effects shown to favour or disfavour the dependent variable are, in fact, shown to favour or disfavour dorsal variants. Four independent variables, or 'factor groups' were included in the analysis (see Table 1): speaker sex, speaker place of origin, preceding phoneme, and following phoneme. With the exception of 'speaker sex', all

[2] The VARBRUL-style analysis was actually implemented in *Rbrul*, operating with different settings. Johnson has shown that, operating in a simpler mode, "Rbrul provided nearly identical output to the actual GoldVarb program" (2009, 381).

factor groups were returned by the regression analysis as significant: place of origin ($p \approx 0$); following phoneme ($p \approx 0$); preceding phoneme ($p \approx 0$). The model presented in Table 3 provides us with the following information: factor groups ('predictors') and factors ('levels'); the total number of tokens considered per factor or level; the percentage of tokens in a given context that are represented by the application value; factor weights; log likelihood; degrees of freedom; significance level; input probability. The '% of total' column reports the percentage of dorsal tokens that occur at each level for a given predictor, e.g., 55% of the tokens that occur before a pause are dorsal in the data set.

Factor weights indicate the "strength and direction of the effect" (Hay 2011, 207); they can range from 0 to 1 and are anchored around the value of 0.5. Factor weights closer to 1 favour the application value, dorsal variants in this case, while factor weights closer to 0 disfavour dorsal variants. Factor weights are important for determining the social and linguistic constraints on the distribution of the variants of the dependent variable. For the (R) variable, dorsal variants, which have been transferred from French, are favoured by speakers from Nay, in pre-pausal and post-pausal position, and following front vowels. It is interesting to note here that dorsal variants are favoured before apical consonants, which may constitute some sort of dissimilatory effect on transfer from French. Apical variants are favoured by speakers from Nousty, before vowels and dorsal consonants, and after all consonant types. These factor weights allow us to interpret the relative strength of the factors across factor groups and, indeed, across sociolinguistic studies.

The log-likelihood value indicates the goodness of fit of the regression model; it is interpreted relative to other log-likelihood values, with more positive values indicating a better fit of the regression model to the data set, or the extent to which the model can explain the variability observed. The degrees of freedom (df) refer to the number of parameters in the model that can vary: "the number of independent pieces of information available or used in the analysis of the observed data" (Paolillo 2002, 109). The degrees of freedom are essentially determined by the number factor groups and factors included in the analysis. All factor groups included in the model are significant at the ($p \approx 0$) level, as indicated by the *p*-value of 0.000. Finally, the 'input' value refers to the input probability. The logistic regression model in Table 3 makes predictions about the proportion of dorsal and apical variants in each cell, given the factor groups included in the analysis; the input probability is basically an average of these predicted values, which "provides the baseline from which the model predictions are built" (Hay 2011, 209).

The stepwise algorithm commonly employed in VARBRUL analyses is now "generally frowned upon in today's statistical community" (Johnson 2009, 380) and outside of the discipline of sociolinguistics, the output presented in Table 3 is difficult to interpret because the terminology (factor groups, factor weights, input probability) are not in use elsewhere. *Rbrul* recommends using a binomial one-level logistic regression model for categorical variables, which takes into account all independent variables considered, including those that do not reach the required level of statistical significance; this has the effect of improving the model's goodness of fit. *Rbrul* has the initial advantage of providing regression output in the form of factor weights, while also presenting the results in log-odds; this facilitates interpretation of the results by sociolinguists that are well-acquainted with VARBRUL, and by researchers in other disciplines where log-odds are more commonly used to interpret the strength and direction of effects for independent variables.

```
ONE-LEVEL ANALYSIS OF RESPONSE Place of Articulation WITH PREDICTOR(S): Following phoneme
(3.13e-06) + Preceding phoneme (0.00019) + Place of origin (0.000223) + Sex (0.184)

$Sex
 factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight
      F   0.184    233                0.275                  0.546
      M  -0.184    233                0.163                  0.454

$Place of origin
 factor    logodds tokens DORSAL/DORSAL+APICAL centered factor weight
    Nay      0.804    145                0.317                  0.691
    Gan      0.405    231                0.221                  0.600
    Nousty  -1.209     90                0.056                  0.230

$Preceding phoneme
         factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight
          Pause   1.945     17                0.412                  0.875
     FrontVowel   0.639    215                0.288                  0.654
      BackVowel   0.239     89                0.281                  0.559
     DorsalCons  -0.109     50                0.100                  0.473
 NonLingualCons  -1.328     61                0.033                  0.209
     ApicalCons  -1.385     34                0.029                  0.200

$Following phoneme
         factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight
          Pause   1.534     66                0.545                  0.823
     ApicalCons   0.293     77                0.273                  0.573
 NonLingualCons   0.038     43                0.256                  0.509
     FrontVowel  -0.392    214                0.121                  0.403
     DorsalCons  -0.709      9                0.111                  0.330
      BackVowel  -0.763     57                0.123                  0.318

$misc.1
   n df intercept overall proportion centered input prob
 466 14    -2.099              0.219               0.109

$misc.2
 log.likelihood     AIC    AICc   Dxy    R2
       -192.545 413.089 414.021 0.628 0.388

Current variables are:
response.binary: Place of Articulation (DORSAL vs. APICAL)
fixed.factor: Sex Place of origin Preceding phoneme Following phoneme
```

Table 4. Rbrul output for one-level simple main effects logistic regression model of Occitan rhotic consonants, with [dorsal] as application value

Table 4 presents the *Rbrul* output for a one-level simple main effects model of the Occitan rhotic consonants, essentially replicating the VARBRUL-style analysis presented in Table 3 without using automated stepwise regression procedures. This logistic regression analysis returns the same predictors are significant: following phoneme (p = 3.13e–06; $p \approx 0$)[3]; preceding phoneme (p = .00019; $p < .001$); place of origin (p = .000223; $p < .001$). Speaker sex is not returned as significant by the model (p = .184), indicating that speaker sex does not have an effect on the choice of dorsal or apical rhotics in this variety of Occitan. The contraint ranking, indicated by the centred factor weights, is identical to the VARBRUL output in Table 3, but the individual factor weight values have changed as a result of the inclusion of 'speaker sex' in the model. While the latter is a non-significant predictor of the variation observed, it nonetheless accounts for some of the variability in the data set and the factor weights for the other predictors are adjusted to reflect this.

Three new pieces of information are provided in the *Rbrul* output in Table 4: the intercept, log-odds values, and an R^2 value. The intercept is similar to the input probability in that it is the baseline value used to built the regression model: "[it] represents the reference levels for all factorial predictors in the model simultaneously" (Tagliamonte/Baayen 2012, 149). *Rbrul* uses 'treatment contrasts' or 'treatment coding' as its default mode of reporting regression output. This means that, for each predictor, one level is chosen as the default; the estimate for the intercept is the mean log-odds for these default values (Tagliamonte/Baayen 2012, 149). Where factor weights are anchored around 0.5 and range on a scale from 0 to 1, log-odds are anchored around zero and range on a scale from negative infinity to positive infinity (Hay 2011, 209; Tagliamonte/Baayen 2012, 148): "we obtain log-odds […] by taking the natural (base *e*) logarithm of the odds, where the odds are the probability of an event occurring, divided by the probability of it not occurring" (Johnson 2009, 361). Positive log-odds favour the application value while negative log-odds disfavour the application value. For example, for the 'following phoneme' predictor, following pauses strongly favour the transfer of dorsal rhotics into Occitan (log-odds = +1.534) while following back vowels strongly favour the retention of apical rhotics (log-odds = –0.763, with [dorsal] as the application value). Values around zero, such as for non-lingual consonants (log-odds = +0.038) have a neutral effect on the dependent variable. An advantage of using the log-odds scale is that the analyst can simply add the log-odds together to obtain an overall prediction for any configuration of the independent variables (Johnson 2009, 361; Hay 2011, 210). For example,

[3] This is an infinitisimal *p*-value, which should be interpreted as 0.000000313 or a decimal point followed by six zeroes and then number before *e*.

if we wanted to predict dorsal consonant use for a speaker from Gan saying a word such as *tribalhar* /trib^{l}ʎa/ where the phonemic apical tap is preceded by an apical consonant and followed by a front vowel, we simply add the log-odds for each of those levels, plus the value of the intercept (Johnson 2009, 361): Gan (0.405) + apical consonant (-1.385) + front vowel (-–0.392) + intercept (–2.099) = –3.471, indicating that this configuration of the independent variables would strongly disfavour the use of a dorsal rhotic. Finally, the R^2 value (0.388) reports the proportion of variation explained by the model, with higher values indicating a more robust explanatory capacity. As with log likelihood, the R^2 value should be interpreted comparatively and relative to the equivalent value in adjusted or modified models of the data set.

5.2 Collinearity and interactions

For the regression analyses of the Occitan rhotic consonants presented in Table 3 (binomial stepwise) and Table 4 (binomial one-level), I noted that the 'syllable type' factor group, or predictor, could not be included in the same model as 'preceding phoneme' and 'following phoneme' because it exhibits a collinear relationship with these predictors; collinearity refers to cases where two or more independent variables are correlated. One of the primary problems with generalised linear models, of which logistic regression is one, is that they assume independent variables not to be collinear. Due to the nature of the wordlist translation task used to collect the Occitan sociophonetic data, the 'syllable type' predictor is correlated with both 'following phoneme' and 'following phonetic environment', since for example, final open syllables (/Cv#/) always correspond to a pause for the 'following phoneme' predictor. In any case, interdependencies such as those between independent predictors should never be considered together (Tagliamonte and Baayen 2012, 163): unsolvable computational problems often arise resulting in various kinds of error message. If the program is successful in returning an output for a model containing collinear predictors, the log-odds estimates for the logistic regression will not be precise because of the difficulty experienced in parcelling out the effect of the collinear predictors on the dependent variable.

In order to include 'syllable type' as an independent variable, 'preceding phoneme' and 'following phoneme' must be excluded from the regression analysis. The resultant one-level simple main effects model is presented in Table 5, where 'syllable type' ($p \approx 0$) and 'place of origin' ($p \approx 0$) are returned as significant effects. The constraints ranking for place

of origin mirrors that of the models in Tables 3 and 4 (including preceding and following phoneme as predictors). For syllable type, dorsal realisations are shown to be favoured in simple syllable codas (log-odds = +1.116), e.g. *mort* [muʁ]), and disfavoured in complex onset clusters (log-odds = –1.182), e.g. *trin* [tr̃ĩ]. Complex codas and simple onsets had a neutral effect on the distribution of dorsal and apical consonants.

```
ONE-LEVEL ANALYSIS OF RESPONSE PlaceArt WITH PREDICTOR(S): Syllable type (1.8e-11) + Place of
origin (0.000327) + Sex (0.253)

$Sex
 factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight
      F   0.151    233                0.275                  0.538
      M  -0.151    233                0.163                  0.462

$Place of origin
 factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight
    Nay   0.778    145                0.317                  0.685
    Gan   0.360    231                0.221                  0.589
 Nousty  -1.138     90                0.056                  0.243

$Syllable type
       factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight
  SimpleCoda   1.116    139                0.417                  0.753
 ComplexCoda   0.088     55                0.200                  0.522
 SimpleOnset  -0.022    116                0.190                  0.494
ComplexOnset  -1.182    156                0.071                  0.235

$misc.1
   n df intercept overall proportion centered input prob
 466  7    -1.779              0.219               0.144

$misc.2
 log.likelihood     AIC   AICc   Dxy    R2
       -204.958 423.916 424.16 0.557 0.299

Current variables are:
response.binary: PlaceArt (DORSAL vs. APICAL)
fixed.factor: Sex Place of origin Syllable type
```

Table 5. Rbrul output for one-level simple main effects logistic regression model (with 'syllable type' as a predictor) of Occitan rhotic consonants, with [dorsal] as application value.

To examine whether it is better to include syllable type as an independent variable in the regression model for the Occitan rhotics, we can compare it with the model including preceding phoneme and following phoneme (see Table 4). The best method for comparing the goodness of fit of two models is to undertake an 'analysis of deviance' using a chi-squared test. The chi-squared test can be performed in *Rbrul* by entering the log-likelihood and degrees of freedom for the first model (Table 4) and the log-likelihood and degrees of freedom for the second model (Table 5); the resultant chi-squared statistic (χ^2) takes into account the difference between the models' log likelihood values and the difference in the number of degrees of freedom. Comparing the models in Tables 4 and 5, the increase in log likelihood from –192.545 (when phonological context is included) to –204.958 (when syllable

type is included) follows a chi-squared distribution ($\chi^2 = 24.362$) with 7 degress of freedom. The *p*-value is .000984 ($p < .001$), indicating that the difference between the two models is significant and that the model in Table 4, containing 'preceding phoneme' and 'following phoneme' as independent variables, is capable of explaining more of the variation observed in the data set.

Johnson (2009, 380) notes that collinearity should not be confused with interactions between independent variables:

> "An *interaction* between two variables exists when the effect of the first variable on the dependent cannot be sensibly discussed or fully represented without reference to the second variable. […] If we are doing multiple regression, we would need to include this interaction in our model, otherwise we might be violating the assumptions of the model, which require that there are no unmodelled interactions in the data" (Hay 2011, 206).

The simple main effects logistic regression model presented in Table 5 considers the predictors, or independent variables, to be completely independent of each other. Simple main effects models ignore potential interactions between independent variables. These interactions are present in many sociolinguistic data sets but have been largely overlooked by VARBRUL-style analyses (cf. Sigley 2003). The exclusion of interactions from variationist analyses is primarily due to the fact that the VARBRUL series does not permit the inclusion of interaction terms in the regression analysis. *Rbrul* allows the inclusion of predictor interaction terms in the "same automatic procedure that identifies significant main effects" (Johnson 2009, 363).

```
ONE-LEVEL ANALYSIS OF RESPONSE PlaceArt WITH PREDICTOR(S): Place of origin (0.000434) +
Sex:Syllable type (0.0199) + Sex [main effect, not tested] + Syllable [main effect, not
tested]

$Sex
 factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight vif
      F   2.059    233                0.275                  0.887 >20
      M  -2.059    233                0.163                  0.113 >20

$Place of origin
 factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight
    Nay   0.774    145                0.317                  0.684
    Gan   0.350    231                0.221                  0.587
 Nousty  -1.124     90                0.056                  0.245

$Syllable type
      factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight vif
  SimpleCoda   2.994    139                0.417                  0.952 >20
 SimpleOnset   1.857    116                0.190                  0.865 >20
ComplexOnset   0.752    156                0.071                   0.68 >20
 ComplexCoda  -5.602     55                0.200                  0.004 >20

$`Sex:Syllable type`
 factor:factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight vif
 F:ComplexCoda   6.158     30                0.367                  0.998 >20
M:ComplexOnset   2.202     76                0.066                    0.9 >20
  M:SimpleCoda   1.992     70                0.343                   0.88 >20
 M:SimpleOnset   1.963     62                0.145                  0.877 >20
 F:SimpleOnset  -1.963     54                0.241                  0.123 >20
  F:SimpleCoda  -1.992     69                0.493                   0.12 >20
F:ComplexOnset  -2.202     80                0.075                    0.1 >20
 M:ComplexCoda  -6.158     25                0.000                  0.002 >20

$misc.1
   n df intercept overall proportion centered input prob
 466 10    -3.643              0.219               0.026

$misc.2
 log.likelihood     AIC    AICc   Dxy   R2
       -200.032 420.064 420.547 0.575 0.81

Current variables are:
response.binary: PlaceArt (DORSAL vs. APICAL)
fixed.factor: Sex Place of origin Syllable
fixed.interaction: Sex:Syllable
```

Table 6. Rbrul output for one-level logistic regression model with interactions of Occitan rhotic consonants, with [dorsal] as application value

Table 6 replicates the logistic regression model in Table 5 for the Occitan rhotic consonants, this time including an interaction term, which considers the relationship between speaker sex and syllable type (sex*syllable). This interaction term is returned by the regression model as significant ($p < .05$), showing that the constraint ranking for syllable type is significantly different for male and female speakers: for female speakers, complex codas favour dorsal variants, with all other contexts favouring apical variants; for male speakers, complex codas favour apical variants, with all other contexts favouring dorsal variants.

The goodness of fit will always be better when interaction terms are included in the regression model "but the improvement is not always worth the complication" (Johnson 2009, 381). To determine if the increase in log likelihood from –204.958 in the simple main effects model (Table 5) to –200.032 in the model with the sex*speaker interaction term (Table 6), we

must perform an analysis of deviance. This increase in log likelihood follows a chi-squared distribution (χ^2 = 9.8519) with 3 degress of freedom ($p < .05$), indicating that this improvement in fit is statistically significant and that the model containing the interaction term should be prioritised.

5.3 Linear regression models

Linear regression is the technique used to model the relationship between a continuous dependent variable, such as a vowel formant frequency, and multiple independent variables (which may be categorical or continuous). This type of analysis is impossible in the VARBRUL series (with the exception of unreleased *Varbrul 3*). It is particularly important, with continuous variables, to know how the data is distributed because the linear regression analysis will assume that is it normally distributed: "in a normal distribution, data are equally distributed around a mean value, and there are two long tails. In such distributions the *mean* (or average) value is approximately the same as the *median* (or middle) value" (Hay 2011, 200). The main problem is that the distribution of many phonetic variables is not normally distributed, in which case it may be advisable to use an alternative, non-parametric test (see Section 6) which makes no assumptions about normality (Hay 2011, 200–201).

The first formant frequency for the Occitan (E) variable was modelled using simple main effects linear regression in *Rbrul* and the results of this analysis are presented in Table 7. The continuous dependent variable is normalised F1; one continuous predictor (F2) and four categorical predictors (speaker sex, place of origin, phoneme /e/ or /ɛ/, syllable type[4]) were included in the analysis. Three predictors were returned as significant: syllable type ($p \approx 0$); phoneme ($p < .01$); F2 ($p < .05$). For the levels of each predictor, regression 'coefficients' are returned by the analysis when the dependent variable is continuous; these coefficients are similar to log-odds in that they indicate the magnitude and direction of the effect identified, with positive values indicating that a level favours a high F1 value (a more open vowel) and negative values favouring a low F1 value (a closer vowel). Coefficients differ from log-odds, however, in that they are expressed in the units of the dependent variable (i.e. F1), rather than on a scale of positive infinity to negative infinity, anchored around zero, as is the case for log-odds (Johnson 2009, 380; Hay 2011, 211).

[4] For reasons of collinearity (see Section 5.2), the 'preceding phoneme' and 'following phoneme' predictors have not been included in the same regression model as 'syllable type'.

```
ONE-LEVEL ANALYSIS OF RESPONSE F1 WITH PREDICTOR(S): Syllable type (3.26e-30) + Phoneme
(0.00179) + F2 (0.0161) + Sex (0.105) + Place of origin (0.635)

$Sex
 factor   coef tokens  mean
      M  0.052    225 0.123
      F -0.052    217 0.066

$Place of origin
 factor   coef tokens  mean
    Nay  0.042    138 0.113
    Gan -0.005    215 0.083
 Nousty -0.037     89 0.096

$Phoneme
 factor   coef tokens   mean
     EH  0.116    246  0.399
      E -0.116    196 -0.287

$Syllable type
    factor   coef tokens   mean
     (_C#)  0.609    144  0.784
 (_CV(C)#) -0.260    140 -0.238
     (C_#) -0.349    158 -0.238

$F2
 continuous  coef
         +1 0.136

$misc.1
   n df intercept overall mean
 442  8    -0.021        0.095

$misc.2
 log.likelihood     AIC    AICc    R2
       -388.059 794.118 794.535 0.414

Current variables are:
response.continuous: F1
fixed.factor: Sex Place of origin Phoneme Syllable
fixed.continuous: F2
```

Table 7. Rbrul output for one-level simple main effects linear regression model of F1; Occitan front mid-vowels.

For syllable type, high F1 values are favoured in final closed syllables (+0.609), while low F1 values are favoured in medial (–0.260) and final (–0.349) open syllables: mid-open vowels occur in closed syllables; mid-close vowels occur in open syllables. For the phoneme predictor, /ɛ/ favours higher F1 values or more open vowels, while /e/ favours lower F1 values or more close vowels, which is the expected pattern. This finding indicates that the phonemic distinction between the Occitan front mid-vowels is maintained for these speakers. Finally, the significant effect for F2 indicates that F2 exhibits a positive correlation with F1 (+0.136): this single coefficient is "the value by which the log odds changes for every increment of 1 in the continuous variable" (Hay 2011, 210). Put simply, for every one unit change in the value of F1, there is a concomitant change of magnitude 0.136 in the value of F2: as the vowel becomes lower, it is also more front; as the vowel becomes higher, it is also more centralised.

A linear regression model of the value of F2 for the (E) variable, including the same independent variables, returned two significant effects: syllable type ($p \approx 0$) and F1 ($p < .05$). A non-significant result for phoneme ($p = .86$) indicated that the phonetic distinction between /e/ and /ɛ/ is not realisaed on the front-back dimension. For syllable type, final open syllables (+0.150) favoured high F2 values, or more front vowels, final closed syllables (–0.238) favoured low F2 values, or more centralised vowels, while medial open syllables (+0.089) had a neutral effect on the value of F2. The coefficient for the F1 predictor (+0.095) revealed the same pattern for F1-F2 correlation as observed in the F1 model (see Table 7): for every unit increase in the value of F2 there is a concomitant increase, of magnitude 0.095, in the value of F1.

5.4 Mixed-effects regression models

One of the assumptions underlying traditional regression analyses is that the individual observations in the data set are independent of each other (Johnson 2009, 363), or that each individual token of the dependent variable is an independent observation. In sociolinguistic data sets, however, tokens are grouped according to the speakers who have produced them: as soon as an individual speaker contributes more than one observation to the data set, they become a source of variation that must be accounted for in the statistical model (Tagliamonte/Baayen 2012, 143). Failure to account for the variation introduced into the data set by individual speakers can lead regression output to grossly overestimate the significance of extralinguistic effects, returning statistically significant results that are likely to be due to individual variation combining with chance (Johnson 2009, 363). Additionally, individual lexical items can potentially favour or disfavour a particular linguistic outcome. It is also necessary to control for this possibility by taking account of variation introduced to the data set by including tokens that have been segmented from a variety of different words. Taking account of variation introduced by differing lexical items leads to more accurate conclusions about internal effects such as phonetic/phonological context or lexical frequency (Johnson 2009, 378).

This section presents mixed-effects logistic and linear regression models for the categorical (R) and continuous (E) Occitan variables. Mixed-effects models make a distinction between two types of predictor that can affect a dependent variable. Firstly, fixed effects are predictors that are replicable in another study, for example, speaker sex

(male/female), stress (tonic/atonic), etc. Random effects, on the other hand, are predictors drawn from a larger population which are not completely replicable (Johnson 2009, 365), such as individual speakers and different lexical items. Including a speaker random effect in the regression analysis takes into account that some individuals may favour a given linguistic outcome while others may disfavour it (Johnson 2009, 365). The mixed model will only return a significant result for a given factor, such as speaker sex, if the effect is strong enough to rise above the inter-speaker variation in the model. Likewise, including lexical item as a random effect takes into account the variation introduced into the model by individual words and only returns a significant result for internal independent variables when their effect is large enough to outweigh inter-lexical-item variation. If individual speakers and individual words are not included in the regression model as random effects, the results of the analysis will only be relevant for the individuals and words sampled and *p*-values may be too small and misleading to generalise to the larger population (Tagliamonte/Baayen 2012, 143).

The VARBRUL series of programs is not optimised for mixed-effects modelling, making it largely unsuitable for the statistical treatment of spontaneous speech data, given the theoretical and methodological advancements made in the field over the past fourty years:

> "As it is usually run, without a factor group for speaker, GoldVarb necessarily ignores the grouping and treats each token as if it were an independent observation. This leads the program to overestimate – potentially drastically – the significance of external effects, those of social factors like gender and age" (Johnson 2009, 363).

VARBRUL can only include individual speaker as a fixed effect in the stepwise regression, often leading to 'speaker' being added as the most significant factor group when stepping up (Johnson 2009, 380), with the resultant model excluding potentially significant factor groups which would be returned as significant by a mixed model. Even if this is not the case, a VARBRUL analysis including speaker as a fixed effect factor group will only be valid for the speakers sampled and we cannot reliably generalise the results to the wider population (Tagliamonte/Baayen 2012, 157). Mixed-effects models take account of individual speakers who particularly favour or disfavour a variant of the dependent variable; the VARBRUL suite will always attempt to attribute disproportionate data of this type to an independent variable or factor group effect whereas a mixed model will "consider holding the speaker responsible" (Johnson 2009, 374). As such, *Rbrul* or direct coding in the R environment offer the best method currently available for the mixed-effects modelling of sociolinguistic data: "the new

mixed-effects models provide the researcher with an even more powerful and principled way of dealing with different kinds of predictors typically encountered in sociolinguistic data sets" (Tagliamonte/Baayen 2012, 143).

5.4.1 Mixed-effects logistic regression

Table 8 presents the *Rbrul* output for a simple one-level mixed-effects logistic regression analysis of the Occitan (R) variable. 'Speaker' and 'word' (lexical item) have been included as random effects; fixed effects included were 'speaker sex', 'place of origin', and 'syllable type'. The regression analysis returned only syllable types as a significant predictor of the distribution of dorsal and apical rhotics ($p \approx 0$): dorsal rhotics are favoured in simple codas; apical rhotics are favoured in complex onsets.

Including 'speaker' as a random effect essentially involves assigning each speaker an individual intercept, or baseline, value, as can be seen in Table 8. This means that the input probability varies between individuals, or that each speakers has a different baseline preference for dorsal variants, and the mixed-model attempts to estimate the magnitude of that preference (Johnson 2009, 381): "this allows speakers to vary randomly with respect to one another, without influencing the overall significance of the investigated effects. The result is that no individual participant can dominate the significance of any reported effect" (Hay 2011, 212). For example, we can see that Speaker B particularly favours the use of dorsal rhotics, while Speaker I particularly disfavours their use; Speakers A and G are relatively neutral with respect to their use of dorsal and apical variants. While random effects do not officially constitute formal parameters in the mixed-model (Johnson 2009, 365), as indicated by [random, not tested] in the first line of the output in Table 8, they behave in a similar way and can be interpreted in the same way as the output for fixed-effect predictors.

Including speaker as a random effect also takes into account the possibility that individuals have different internal constraints on variation (Johnson 2009, 374). Incorporating this information into our analyses allows us to address the much-debated 'homogeneity assumption', that is that speakers share a common grammar with common constraints on variation: "this helps us avoid the criticism that individual speaker agency is lost in quantitative analyses using social categories" (Johnson 2009, 377). Equally, including lexical item as a random effect (estimates not shown in Table 8), accounts for the possibility that individual words can favour or disfavour a dorsal variant due to 'word specific phonetics' (cf. Pierrehumbert 2006).

Comparing this mixed-effects regression with the simple main effects logistic regression presented in Table 5, where speaker and word are not included as random effects, we can first note that the goodness of fit is significantly better for the mixed model (χ^2 = 24.172; df = 2; p = 5.64e–06). Additionally, 'place of origin' was included as a significant predictor in the simple main effects model ($p \approx 0$) (see Table 5), but this predictor is returned as non-significant by the mixed-effects analysis. When there is a lot of individual speaker variation in the data set, variation in the dependent variable that is occurring by chance can be wrongly interpreted by a simple logistic model as correlated to an independent variable: "change can create the appearance of external effects" (Johnson 2009, 365). Mixed-effects modelling in *Rbrul* reduces this possibility by only returning a significant result for independent variables when their effect is strong enough to rise above the variation between individual speakers. This conservative approach often leads to less significant results in mixed-effects models, such as the exclusion of the 'place of origin' predictor, but we can be much more confident about the effects that are returned as significant (Hay 2011, 212–213).

```
ONE-LEVEL ANALYSIS OF RESPONSE PlaceArt WITH PREDICTOR(S): Speaker [random, not tested] and
Word [random, not tested] and Syllable type (7.9e-08) + Place of origin (0.277) + Sex (0.357)

$Sex
 factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight
      F   0.349    233                0.275                  0.586
      M  -0.349    233                0.163                  0.414

$Place of origin
 factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight
    Nay   0.878    145                0.317                  0.706
    Gan   0.205    231                0.221                  0.551
 Nousty  -1.083     90                0.056                  0.253

$Syllable type
      factor logodds tokens DORSAL/DORSAL+APICAL centered factor weight
  SimpleCoda   1.320    139                0.417                  0.789
 SimpleOnset   0.068    116                0.190                  0.517
 ComplexCoda   0.023     55                0.200                  0.506
ComplexOnset  -1.412    156                0.071                  0.196

$`Word (random) [...]

$`Speaker (random)`
         intercept tokens DORSAL/DORSAL+APICAL centered factor weight
 std dev     0.925    466                0.219                    ...
     ...       ...    ...                  ...                    ...
       B     1.405     47                0.447                  0.789
       J     0.942     44                0.295                  0.702
       H     0.822     51                0.392                  0.677
       F     0.312     45                0.067                  0.557
       G    -0.009     45                0.044                  0.477
       A    -0.037     49                0.204                   0.47
       D     -0.29     44                0.295                  0.408
       E    -0.384     43                0.163                  0.385
       C    -0.399     50                 0.26                  0.382
       I    -1.521     48                    0                  0.167

$misc.1
   n df intercept overall proportion centered input prob
 466  9      -2.1              0.219               0.109

$misc.2
 log.likelihood     AIC    AICc Dxy.fixed Dxy.total R2.fixed R2.random R2.total
       -192.872 403.744 404.139         1     0.761    0.308     0.197    0.505

Current variables are:
response.binary: PlaceArt (DORSAL vs. APICAL)
fixed.factor: Sex Place of origin Syllable type
random.intercept: Speaker Word
```

Table 8. Rbrul output for one-level mixed-effects logistic regression model of Occitan rhotic consonants, with [dorsal] as application value.

Rbrul is also capable of running mixed-effects logistic regression models with interaction terms or, put simply, a mixed-effects version of the regression analysis in Table 6. This also results in a significant improvement in goodness of fit (χ^2 = 27.912; df = 11; p = .00334) and in 'place of origin' being excluded as a significant predictor of the transfer of dorsal rhotics from French (p = .374 versus p = .000434 in Table 6). Including a speaker sex*syllable type interaction term in the mixed-effects analysis does not, however, result in a significant improvement over the simple mixed-effects model presented in Table 8 (χ^2 =

13.592; df = 12; p = .328); in short, including the interaction term is not 'worth the complication' in this case.

5.4.2 Mixed-effects linear regression

Table 9 presents the mixed-effects linear regression analysis of F1 for the Occitan (E) variable, including speaker and word as random effects. Three predictors are returned as significant: syllable type ($p \approx 0$); phoneme ($p < .01$); F2 ($p < .01$). The strength and magnitude of the effects are comparable to those returned by the simple linear regression model presented for the same data in Table 7, where speaker and word were not considered as random effects; the constraints ranking is also identical for each predictor, though the individual coefficient values for each level vary slightly. The improvement in log likelihood from –388.059 (Table 7) to –376.342 (Table 9), when random effects are included, is significant, following a chi-squared distribution ($\chi^2 = 23.434$) with 3 degrees of freedom and probability p = 3.28e–05. Again, mixed-effects modelling can explain more of the variation present in the data set and provides us with more accurate predictions for the independent variables included in the analysis.

```
ONE-LEVEL ANALYSIS OF RESPONSE F1 WITH PREDICTOR(S): Speaker.ID [random, not tested] and Word
[random, not tested] and Syllable type (1.63e-17) + Phoneme (0.0045) + F2 (0.00466) + Sex
(0.197) + Place of origin (0.718)

$Sex
 factor   coef tokens  mean
      M  0.045    225 0.123
      F -0.045    217 0.066

$Place of origin
 factor   coef tokens  mean
    Nay  0.035    138 0.113
    Gan  0.004    215 0.083
 Nousty -0.039     89 0.096

$Phoneme
 factor  coef tokens   mean
     EH  0.11    246  0.399
      E -0.11    196 -0.287

$Syllable type
    factor   coef tokens   mean
     (_C#)  0.552    144  0.784
 (_CV(C)#) -0.220    140 -0.238
     (C_#) -0.332    158 -0.238

$F2
 continuous  coef
         +1 0.156

$`Word (random)[...]

$`Speaker (random)[...]

$misc.1
   n df intercept overall mean
 442 11    -0.017        0.095

$misc.2
 log.likelihood     AIC   AICc R2.fixed R2.random R2.total
       -376.342 809.726 810.34    0.369     0.149    0.518


Current variables are:
response.continuous: F1
fixed.factor: Sex Place of origin Phoneme Syllable
fixed.continuous: F2
random.intercept: Speaker.ID Word
```

Table 9. Rbrul output for one-level mixed-effects linear regression model of F1; Occitan front mid-vowels.

For F2, an equivalent model returned only syllable type ($p \approx 0$) and F1 ($p < .05$) as significant predictors of the variation obeserved; F2 is not used to realise phonetically the phonemic distinction between /e/ and /ɛ/. Including speaker and word as random effects does not, however, result in a significant improvement when compared with a simple main effects linear regression, without random effects ($\chi^2 = .5380$; df = 3; $p = .91$). This may be because F2 is less important for distinguishing between the front mid-vowels, which are traditionally distinguished on the vowel height dimension in articulatory terms. The vocalic data has also been normalised across speakers before both statistical analyses, reducing the amount of inter-speaker variation in the data set, even when 'speaker' is not included as a random effect.

6. Other quantitative approaches: non-parametric tests

As previously noted, regression assumes that the data used in the analyses are normally distributed. Furthermore, mixed-effects modelling also assumes the normality of random effects (Johnson 2009, 381). The unbalanced nature of sociolinguistic data sets, however, often means that the data is not normally distributed, thus posing problems regarding the reliability of the results obtained from the regression output. One possible solution to this problem is submitting the data to a non-parametric statistical analysis, such as random forests, which make no assumptions about the underlying distribution of the data set (Baayen 2008, 77). These methods have the additional advantage of being able to consider independent variables that exhibit a collinear relationship with each other which, as we have seen, can pose serious problems for regression modelling (see Section 5.2).

Random forests (Breiman 2001) are a form of CART (classification and regression trees) analysis, a non-parametric clustering technique used to model non-normal data: "like logistic models, random forests seek to predict, given a set of predictors, which of the alternatives [...] is most probable. However, these statistical techniques achieve the same goal quite differently" (Tagliamonte/Baayen 2012, 159). This technique allows many (potentially correlated) predictors to be considered in the analysis and gives a precise ranking of independent variable importance (Hay 2011, 213). Random forests aim to determine whether an independent variable is a significant predictor by dividing the data set into subsets; individual 'conditional inference trees' are constructed for each independent variable on the basis of these subsets, which estimate the likelihood of a given value for the dependent variable (Tagliamonte/Baayen 2012, 159). When conditional inference trees have been constructed, each tree receives a 'vote' regarding the most likely outcome for the dependent variable, which feeds into the final output for the model (Hay 2011, 213). The 'party' package in the R environment makes it possible to construct forests of these conditional inference trees, but this technique is markedly more time-consuming that regression modelling (Tagliamonte/Baayen 2012, 165).

7. Conclusion

Sociolinguistic theory and methods, including quantitative statistical analysis, have traditionally developed in a predominantly anglophone context, leading to a dearth of equivalent literature for the Romance languages. This is particularly true of Gallo-Romance

with the exception, perhaps, of French in Canada. Additionally, minority languages in the francophone context have received little or no attention in the variationist sociolinguistic literature. The application of quantitative statistical modelling techniques to data from Occitan, presented in this chapter, had the concomitant aims of providing an overview of the evolution of statistical practice in sociolinguistics and of providing an empirical contribution to the field of Occitan dialectology. Beginning with the traditional sociolinguistic concept of the variable rule, the Occitan rhotic consonants were submitted to a VARBRUL-style logistic regression analysis using an automated stepwise algorithm which has since been discredited in general statistical practice: "if VARBRUL and its successor GoldVarb were cutting-edge when introduced, this is no longer the case" (Johnson 2009, 376). More recent statistical practice recommends the use of a one-level generalised linear model for logistic regression, which takes into account the influence of non-significant predictors on the distribution of the variation observed. One problem with this technique is that independent variables that are heavily correlated cannot be included in the same regression analysis, but these correlations should not be confused with interactions between independent variables. The inclusion of interaction terms, facilitated by the *Rbrul* package, can significantly improve the fit of a regression model, as demonstrated using an analysis of deviance test for the Occitan rhotic consonants. An additional advantage of using *Rbrul* was demonstrated in the analysis of the first two formant frequencies of the Occitan front mid-vowels using a linear regression analysis for continuous dependent variables, previously impossible using the VARBRUL series. The most important advancement in statistical practice for sociolinguistics has been the introduction of generalised mixed-effects models, which make it possible to model subtle differences among internal and external independent variables (Tagliamonte/Baayen 2012, 136), by considering the variability introduced into the data set by individual speakers and lexical items before calculating the direction and magnitude of effects for independent variables. For both categorical and continuous dependent variables, the analysis of the Occitan data showed that, in the majority of cases, mixed-effects regression results in an improved predictive model for the unbalanced data sets used in sociolinguistic studies: "mixed-model analysis using *Rbrul* giver better estimates not only for the significance (*p*-values) of external effects, but for their sizes (coefficients), as well" (Johnson 2009, 378). Nonetheless, issues such as collinearity continue to pose serious problems for mixed-effects modelling, with non-parametric techniques, such as random forests, providing a useful, if time-consuming, alternative to regression analysis.

8. Bibliography

Adank, Patti/Smits, Roel/van Hout, Roeland (2004), *A comparison of vowel normalisation procedures for language variation research*, Journal of the Acoustical Society of America 116, 3099–3107.

Agresti, Alan (2007), *An introduction to categorical data analysis*, Hoboken, NJ: Wiley.

Ashby, William J. (1981), *The loss of the negative particle ne in French: A syntactic change in progress*, Language 57, 674–687.

Ashby, William J. (1982), *The drift of French syntax*, Lingua 57, 29–46.

Ashby, William J. (1988), *The syntax, pragmatics, and sociolinguistics of left-and right-dislocations in French*, Lingua 75, 203–229.

Ashby, William J. (1992), *The variable use on versus tu/vous for indefinite reference in Spoken French*, Journal of French Language Studies 2, 135–157.

Bates, Douglas/Sarkar, Deepayan (2008), *lme4: linear mixed-effects models using S4 classes*, <http://cran.r-project.org/> (05.08.2016)

Bec, Pierre (1973), *Manuel pratique d'occitan moderne*, Paris: A & J Picard.

Blanchet, Philippe/Schiffman, Harold (2004), *Revisiting the sociolinguistics of "Occitan": a presentation*, International Journal of the Sociology of Language 169, 3–24.

Boersma, Paul (2001), *Praat: A system for doing phonetics by computer*, Glot International 5, 341–345.

Boersma, Paul/Weenink, David (2012), *Praat: Doing phonetics by computer*, <http://www.praat.org/> (05.08.2016).

Breiman, Leo (2001), *Random forests*, Machine Learning 45, 5–32.

Burnett, Heather/Tremblay, Mireille/Blondeau, Hélène (2015), *The Variable Grammar of Negative Concord in Montréal French*, University of Pennsylvania Working Papers in Linguistics 21, 10–20.

Cardaillac Kelly, Reine (1973), *A Descriptive Analysis of Gascon*, The Hague: Mouton.

Cedergren, Henrietta J./Sankoff, David (1974), *Variable rules: performance as a statistical reflection of competence*, Language 50, 333–55.

Chen, Yiya (2008), *The acoustic realization of vowels of Shanghai Chinese*, Journal of Phonetics 36, 629–48.

Clopper, Cynthia. G. (2011), *Checking for reliability*, in Marianna Di Paolo/Malcah Yaeger-Dror (edd.), *Sociophonetics: a student's guide*, Abingdon: Routledge, 188–197.

Comeau, Philip/King, Ruth/Butler, Gary R. (2012), *New insights on an old rivalry: The passé simple and the passé composé in spoken Acadian French*, Journal of French Language Studies 22, 315–343.

Fabricius, Anne. H./Watt, Dominic/Johnson, Daniel E. (2009), *A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics*, Language Variation and Change 21, 413–435.

Flynn, Nicholas (2011), *Comparing vowel formant normalisation procedures*, York Working Papers in Linguistics 11, 1–28.

Flynn, Nicholas/Foulkes, Paul (2011), *Comparing Vowel Formant Normalization Methods*, Proceedings of the 17th ICPhS, 683–686.

Fougeron, Cécile/Smith, Caroline L. (1999), *Illustrations of the IPA: French*, Handbook of the International Phonetic Association, Cambridge: Cambridge University Press, 78–81.

Hay, Jennifer. 2011. *Statistical analysis*, in Marianna Di Paolo/Malcah Yaeger-Dror (edd.), *Sociophonetics: a student's guide*, Abingdon: Routledge, 198–214.

Johnson 2009. *Getting off the Goldvarb standard: Introducing Rbrul for mixed-effects variable rule analysis*, Language and Linguistics Compass 3, 359–383.

Kennard, Holly J./Lahiri, Aditi (2015a), *Mutation in Breton verbs: pertinacity across generations*, Journal of Linguistics, Advance online publication. doi:10.1017/S0022226715000420

Kennard, Holly J./Lahiri, Aditi (2015b), *Maintenance of the Breton mixed mutation*, Proceedings of the 18th ICPhS, Paper 607.

King, Ruth/Nadasdi, Terry (2003), *Back to the future in Acadian French*, Journal of French Language Studies 13, 323–337.

King, Ruth/Martineau, France/Mougeon, Raymond (2011), *The interplay of internal and external factors in grammatical change: First-person plural pronouns in French*, Language 87, 470–509.

Labov, William (1969), *Contraction, deletion and inherent variability of the English copula*, Language 45, 715–762.

Labov. William (2007), *Transmission and diffusion*, Language 83, 344–387.

Llamas, Carmen/Watt, Dominic/Johnson, Daniel E. (2009), *Linguistic Accommodation and Salience of National Identity Markers in a Border Town*, Journal of Language and Social Psychology 28, 381–407.

Lobanov, Boris M (1971), *Classification of Russian vowels spoken by different speakers*, Journal of the Acoustical Society of America 49, 606–608.

Moisset, Christine (2000), *Variable liaison in parisian French*, unpublished doctoral thesis, University of Pennsylvania.

Mooney, Damien (2014), *Illustrations of the IPA: Béarnais (Gascon)*, Journal of the International Phonetic Association 44, 343–350.

Mooney, Damien (2015a), *Confrontation and language policy: non-militant perspectives on conflicting revitalisation strategies in Béarn, France*, in Mari C. Jones (ed.), *Policy and Planning for Endangered Languages*, Cambridge, Cambridge University Press, 153–170.

Mooney, Damien (2015b) *Transmission and diffusion: Linguistic change in the regional French of Béarn*, Journal of French Language Studies, Advance online publication. doi.org/10.1017/S0959269515000290.

Mooney, Damien (2016a). *'C'est jeuli, la Gasceugne!' L'antériorisation du phonème /ɔ/ dans le français régional du Béarn*, French Studies 70: 61–81.

Mooney, Damien. 2016b. *Southern Regional French: A Linguistic Analysis of Language and Dialect contact*. Oxford: Legenda.

Moreux, Bernard (2004), *Béarnais and Gascon today: language behaviour and perception*, International Journal of the Sociology of Language 169, 25–62.

Nagy, Naomi/Blondeau, Hélène (1999), *Double subject marking in L2 Montreal French*, University of Pennsylvania Working Papers in Linguistics 6, 93–108.

Nearey, Terrance M. (1977), *Phonetic Feature Systems for Vowels*, unpublished doctoral thesis, University of Alberta.

Paradis, Claude/Deshaies, Denise (1990), *Rules of stress assignment in Québec French: Evidence from perceptual data*, Language Variation and Change 2, 135–154.

Paolillo, John C. (2002a), *Analyzing linguistic variation: statistical models and methods*, Stanford, CA: CSLI Publications.

Pierrehumbert, Janet B. (2006), *The next toolkit*, Journal of Phonetics 34, 516–530.

Poplack, Shana. (1991), *The inherent variability of the French subjunctive*, in Christiane Laeufer/Terrell A. Morgan (edd.), *Theoretical analysis in Romance linguistics*, Amsterdam, John Benjamins, 235–263.

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, Vienna, R Foundation for Statistical Computing, <http://www.R-project.org.> (05.08.2016).

Rand, David/Sankoff, David (1990), *GoldVarb: A variable rule application for the Macintosh*, Montréal, Centre de recherches mathématiques, Université de Montréal.

Regan, Vera (1996), *Variation in French interlanguage: A longitudinal study of sociolinguistic competence*, Amsterdam, John Benjamins.

Roberts, Nicholas. S. (2012), *Future temporal reference in Hexagonal French*, University of Pennsylvania Working Papers in Linguistics 18, 97–106.

Rousseau, Pascale/Sankoff, David (1978a), *Advances in variable rule methodology*, in David Sankoff (ed.), *Linguistic variation: models and methods*, New York, NY: Academic Press, 57–69.

Sankoff, David (1988), *Sociolinguistics and syntactic variation*, in Frederick H. Newmeyer (ed.), *Linguistics: the Cambridge Survey: Volume IV. Language: The socio-cultural context*, Cambridge, Cambridge University Press, 140–161.

Sankoff, David. (2006), *Variable rules*, in Ulrich Ammon (ed.), *Sociolinguistics: an international handbook of the science of language and society*, Berlin, Walter de Gruyter, 1150–1163.

Sankoff, David/Labov, William (1979), *On the uses of variable rules*, Language in Society 8, 189–222.

Sankoff, David/Tagliamonte, Sali/Smith, Eric (2005). *GoldVarb X: a variable rule application for Macintosh and Windows*, University of Toronto.

Sankoff, Gillian (2005), *Cross-sectional and longitudinal studies in sociolinguistics*, in Ulrich Ammoon/Norbert Dittmar/Klaus J. Mattheier/Peter Trudgill (edd.), *International handbook of the science of language and society*, Berlin, Mouton de Gruyter, 1003–1013.

Sankoff, Gillian/Blondeau, Hélène (2007), *Language change across the lifespan: /r/in Montreal French*, Language 83, 560–588.

Séguy, Jean (1954–1973), *Atlas Linguistique et Ethnographique de la Gascogne*, Paris: CNRS.

Sichel-Bazin, Rafèu/Buthke, Carolin/Meisenburg, Trudel (2012), *The prosody of Occitan-French bilinguals*, in Kurt Braunmüller/Christoph Gabriel (edd.), *Multilingual individuals and multilingual societies*, Amsterdam, John Benjamins, 349–364.

Sigley, Robert (2003), *The importance of interaction effects*, Language variation and change 15, 227–253.

Sykes, Alan O. (1993), *An introduction to regression analysis*, Coase-Sandor Institute for Law and Economics Working Paper 20, 1–33.

Tagliamonte, Sali (2006), *Analysing Sociolinguistic Variation*, Cambridge, Cambridge University Press.

Tagliamonte, Sali/Baayen, R. Harald (2012), *Models, forests and trees of York English: Was/were variation as a case study for statistical practice*, Language Variation and Change 24, 135–178.

Temple, Rosalind A. (2000a), *Now and then: The evolution of male-female differences in the voicing of consonants in two varieties of French*, Leeds Working Papers in Linguistics and Phonetics 8, 193–204.

Temple, Rosalind A. (2000b), *Old wine into new wineskins. A variationist investigation into patterns of voicing in plosives in the Atlas Linguistique de la France*, Transactions of the Philological Society 98, 353–394.

Thomas, Eric R/Kendall, Tyler (2007). *NORM: The vowel normalization and plotting suite*, < http://lingtools.uoregon.edu/norm/norm1.php> (05.08.2016)

Van Compernolle, Rémi A. (2008), *Morphosyntactic and phonological constraints on negative particle variation in French-language chat discourse*, Language Variation and Change 20, 317–339.

Villeneuve, Anne-José/Auger, Julie (2013) *'Chtileu qu'i m'freumereu m'bouque i n'est point coér au monne': Grammatical variation and diglossia in Picardie*, Journal of French Language Studies 23, 109–133.

Walter, Henriette (1988), *Le francais dans tous les sens*, Paris, Robert Laffont.

Watt, Dominic/Fabricius, Anne (2002), *Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1–F2 plane*, Leeds Working Papers in Linguistics and Phonetics 9, 159–173.

Weinreich, Uriel/Labov, William/Herzog, Marvin I. (1968), *Empirical foundations for a theory of language change*, in Winfred P. Lehmann/Yakov Malkiel (edd.), *Directions for historical linguistics: a symposium*, Austin, TX, University of Texas Press, 95–195.

Williams, Lawrence/van Compernolle, Rémi A. (2009), *On versus tu and vous: Pronouns with indefinite reference in synchronous electronic French discourse*, Language Sciences 31, 409–427.